Прикладная статистика 8. Регрессия бинарного признака.

Рябенко Евгений
riabenko.e@gmail.com

7 апреля 2014 г.

Постановка

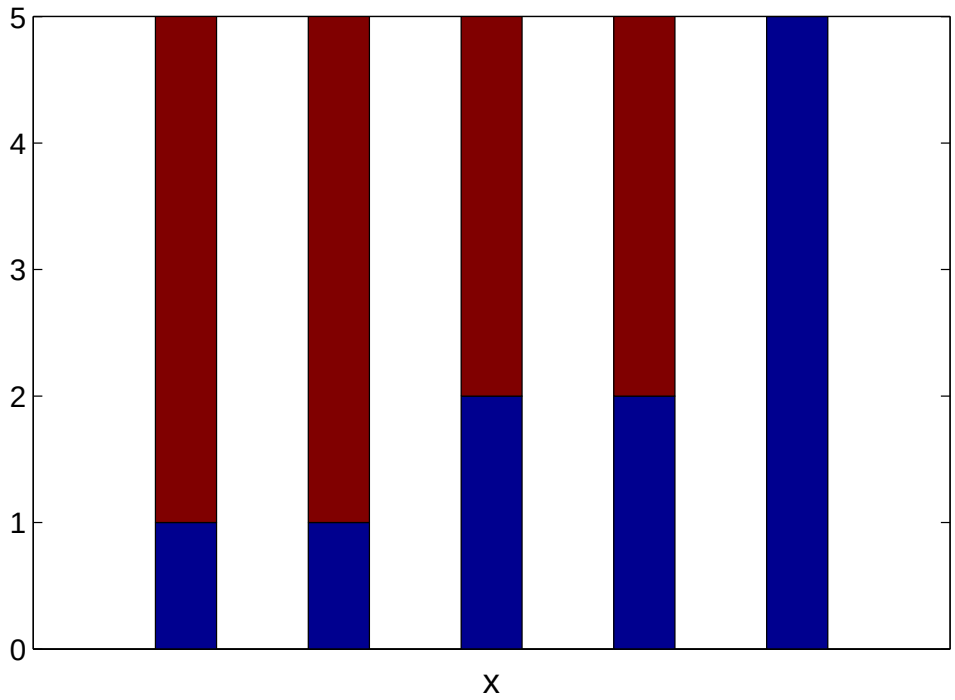
Задача: оценить влияние одного или нескольких признаков на наступление какого-либо события и оценить его вероятность.

$1, \ldots, n$ — объекты;
$x_1, \ldots, x_k$ — предикторы;
y — отклик, $y_i \in \{0, 1\}$.

$$P\left(y = 1 \,|x\right) \equiv \pi\left(x\right) = ?$$

Пример 1

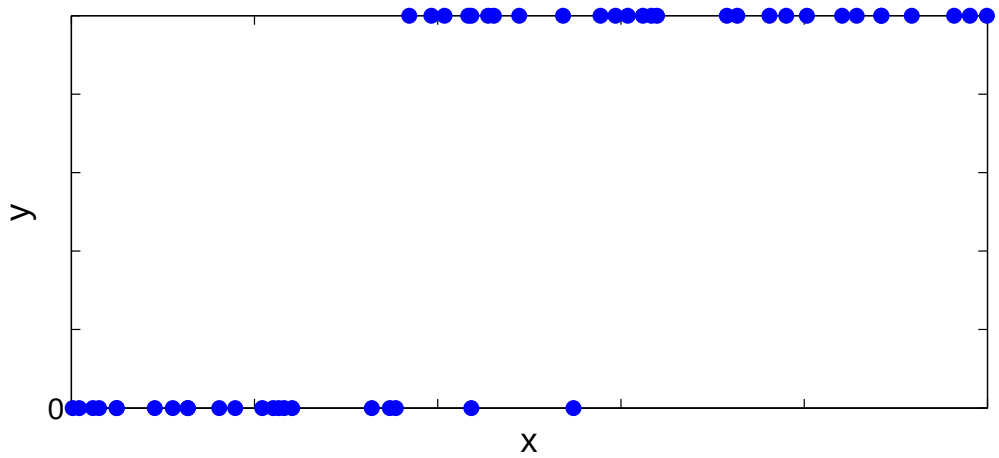
Разработка пестицидов: x_i — доза пестицида, y_i — смерть вредителя.
Повторяемый эксперимент с фиксированными уровнями фактора:



$$\hat{\pi}(x) = \frac{\sum_{i=1}^{n} y_i [x = x_i]}{\sum_{i=1}^{n} [x = x_i]}.$$

Пример 2

Построение кривой спроса: x_i — цена товара, y_i — согласие купить товар.
Неповторяемый эксперимент со случайными уровнями фактора:



Можно построить непараметрическую оценку при помощи ядерного сглаживания:

$$\hat{\pi}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}.$$

Параметризация

Линейная регрессия:

$$\pi\left(x\right) = \beta_0 + \beta_1 x + \varepsilon.$$

- Оценка вероятности может выходить за $[0, 1]$.
- В линейной регрессии $y = \mathbb{E}\left(y\,|x\right) + \varepsilon$, и МНК-оценка β хороша, когда $\varepsilon \sim N\left(0, \sigma\right)$. Здесь же, если $y = \pi\left(x\right) + \varepsilon$, то $\varepsilon = 1 - \pi(x)$ или $\varepsilon = \pi(x)$, и МНК-оценка будет плохой.

Нужно такое нелинейное преобразование

$$g\left(\pi\left(x\right)\right) = \beta_0 + \beta_1 x + \varepsilon,$$

чтобы:

- $\hat{\pi}\left(x\right) = g^{-1}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right)$ принимала значения из $[0, 1]$;
- изменения на краях диапазона значений x приводили к меньшим изменениям $\pi(x)$:
  x — годовой доход, y — покупка автомобиля,

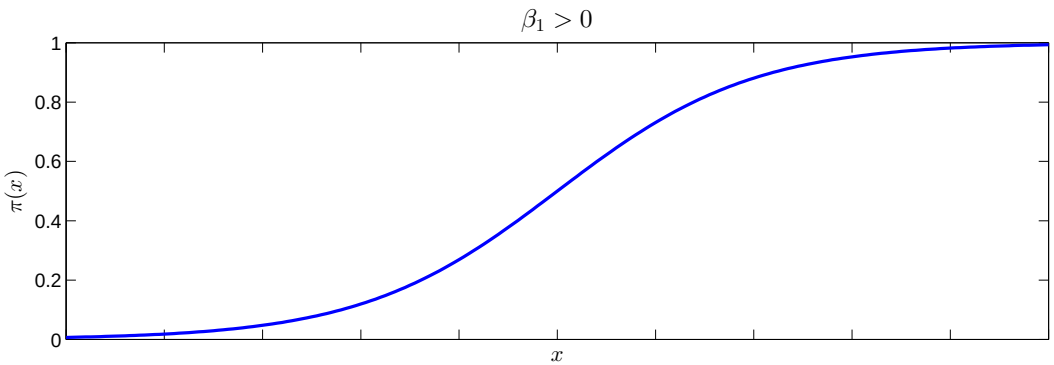
$$\pi\left(10000000 + 200000\right) - \pi\left(10000000\right) < \pi\left(500000 + 200000\right) - \pi\left(500000\right).$$

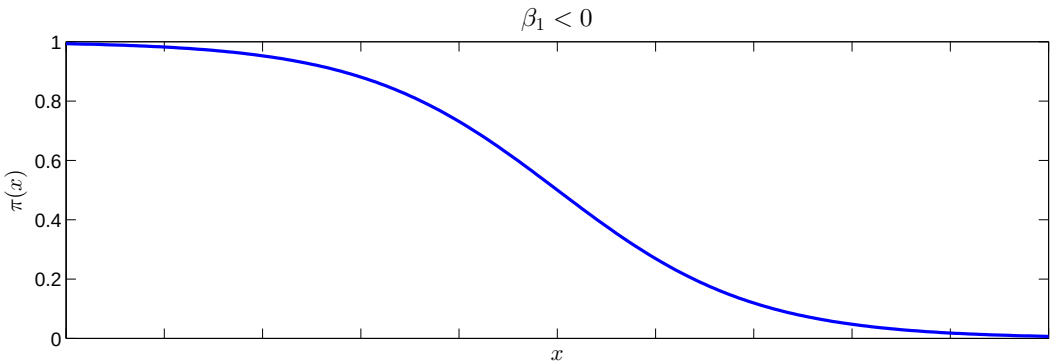
Параметризация

Logit:

$$g(x) = g(\pi(x)) = \ln \frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1 x + \varepsilon,$$

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1+e^{\hat{g}(x)}} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}}.$$





Относительный риск

Пусть $y \sim Ber(p)$, тогда **риск (odds)** события $y = 1$:

$$ODDS = \frac{p}{1-p}.$$

Если $y_1 \sim Ber(p_1)$, $y_2 \sim Ber(p_2)$, то **относительный риск (odds ratio)** события $y_1 = 1$ по сравнению с событием $y_2 = 1$:

$$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$

Серд. заболевания \ Возраст	≥ 55	≤ 55
есть	21	22
нет	6	51

$$OR = \frac{21/6}{22/51} \approx 8.1.$$

Роль коэффициентов регрессии

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Пусть $x = [\text{возраст} \geq 55]$, $y = [\text{есть сердечные заболевания}]$. По $\hat{\beta}_1$ легко оценить относительный риск получения заболевания пожилыми людьми:

$$\widehat{OR} = e^{\hat{\beta}_1}.$$

Пусть $x = \text{возраст}$, $y = [\text{есть сердечные заболевания}]$. $e^{\hat{\beta}_1}$ имеет смысл мультипликативного прироста риска получения заболевания при увеличении возраста на 1 год.

Настройка параметров

$\pi(x)$ оценивает $P(y=1|x)$,
$1-\pi(x)$ оценивает $P(y=0|x) \Rightarrow$

$$P(x_i, 1) = \pi(x_i),$$
$$P(x_i, 0) = 1 - \pi(x_i),$$
$$l(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i} (1 - \pi(x_i))^{1-y_i},$$
$$L(\beta) = \ln l(\beta) = \sum_{i=1}^{n} (y_i \ln \pi(x_i) + (1-y_i) \ln(1 - \pi(x_i))),$$
$$\hat{\beta} = \operatorname{argmax}_{\beta} L(\beta).$$

$\hat{\beta}$:

- существует и единственна,
- находится методом Ньютона-Рафсона,
- является состоятельной и асимптотически эффективной оценкой β,
- асимптотически нормальна.

Проблемы МП-оценки

$\hat{\beta}$ может не существовать или не быть конечной, если:

- наблюдения $y = 0$ и $y = 1$ линейно разделимы в пространстве признаков X;
- матрица X вырождена.

Итерационный процесс может не сойтись, если число признаков k слишком велико относительно числа наблюдений n.

Дисперсия оценок

Пусть $I(\beta) \in \mathbb{R}^{(k+1)\times(k+1)}$ — матрица вторых производных $L(\beta)$:

$$\frac{\partial^2 L}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 \pi(x_i)(1-\pi(x_i)),$$

$$\frac{\partial^2 L}{\partial \beta_j \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} \pi(x_i)(1-\pi(x_i)).$$

Другая форма записи:

$$I(\beta) = X^T V X,$$

$$V = \operatorname{diag}\left(\pi(x_1)(1-\pi(x_1)), \ldots, \pi(x_n)(1-\pi(x_n))\right).$$

Из теории оценок максимума правдоподобия: $\mathbb{D}\hat{\beta} = I^{-1}\left(\hat{\beta}\right)$.

Доверительные интервалы

Для отдельного коэффициента β_j:

$$\hat{\beta}_j \pm z_{1-\alpha/2}\sqrt{\left(I^{-1}\left(\hat{\beta}\right)\right)_{jj}}.$$

Для $g(x_0)$ — логита нового объекта x_0:

$$x_0^T\hat{\beta} \pm z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}.$$

Для вероятности $y = 1$ при $x = x_0$:

$$\left[\frac{e^{x_0\hat{\beta}-z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}{1+e^{x_0\hat{\beta}-z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}, \frac{e^{x_0\hat{\beta}+z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}{1+e^{x_0\hat{\beta}+z_{1-\alpha/2}\sqrt{x_0^T I^{-1}\left(\hat{\beta}\right)x_0}}}\right].$$

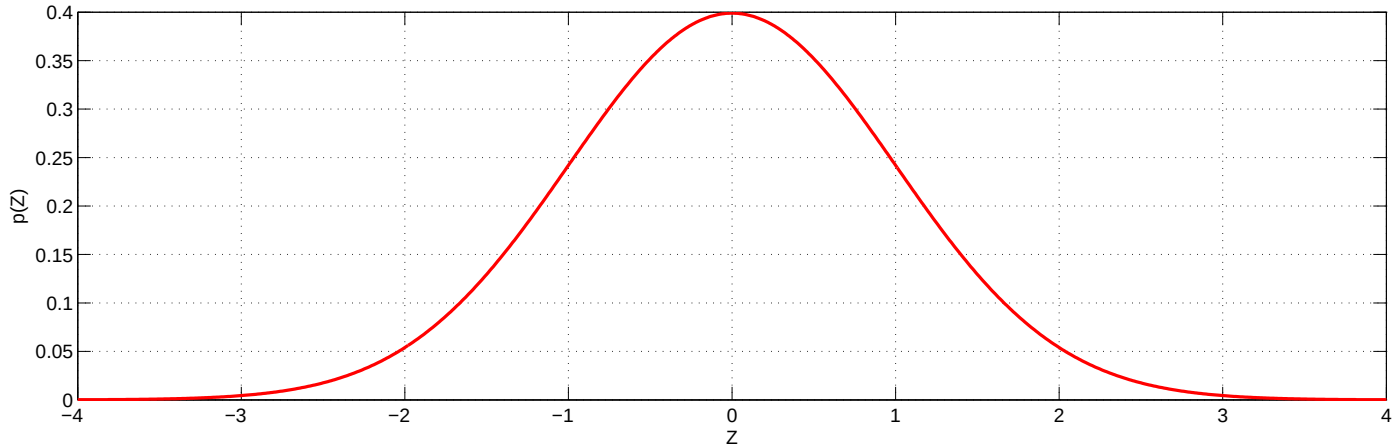
Критерий Вальда

нулевая гипотеза: $H_0 : \beta_j = 0;$

альтернатива: $H_1 : \beta_j <\neq> 0;$

статистика: $T = \dfrac{\hat{\beta}_j}{\sqrt{\left(I^{-1}(\hat{\beta})\right)_{jj}}};$

$T \sim N(0,1)$ при $H_0;$



достигаемый уровень значимости:

$$p(t) = \begin{cases} 1 - ncdf(t,0,1), & H_1 : \beta_j > 0, \\ ncdf(t,0,1), & H_1 : \beta_j < 0, \\ 2\left(1 - ncdf(|t|,0,1)\right), & H_1 : \beta_j \neq 0. \end{cases}$$

Критерий отношения правдоподобия

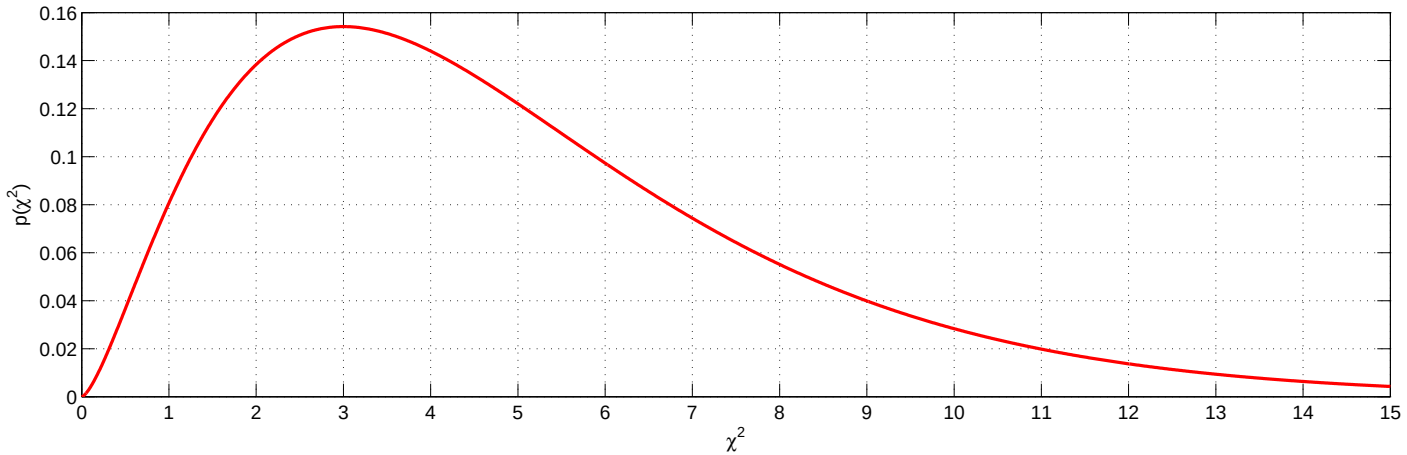
$$\underset{n\times(k+1)}{X} = \left(\underset{n\times(k+1-k_1)}{X_1}, \underset{n\times k_1}{X_2}\right); \qquad \underset{(k+1)\times 1}{\beta^T} = \left(\underset{(k+1-k_1)\times 1}{\beta_1^T}, \underset{k_1\times 1}{\beta_2^T}\right)^T;$$

нулевая гипотеза: $H_0 : \beta_2 = 0;$

альтернатива: $H_1 : H_0$ неверна;

статистика: $G = 2\left(L_r - L_{ur}\right);$

$G \sim \chi^2_{k_1}$ при $H_0;$



достигаемый уровень значимости:

$$p\left(g\right) = 1 - chi2cdf\left(g, k_1\right).$$

Связь между критериями Вальда и отношения правдоподобия

При $k_1 = 1$ критерии Вальда и отношения правдоподобия не эквивалентны, в отличие от случая линейной регрессии, когда в этом случае достигаемые уровни значимости критериев Стьюдента и Фишера совпадают.

При больших n и $\sum_{i=1}^{n} [y_i = 1]$ разница между критериями невелика, но в случае, когда их показания расходятся, рекомендуется смотреть на результат критерия отношения правдоподобия.

Значимость категориальных предикторов

Значимость фиктивных переменных, кодирующих один категориальный предиктор, — тонкий вопрос.

- Необходимо включать или исключать категориальный предиктор целиком. Значимость соответствующих фиктивных переменных проверяется в совокупности с помощью критерия отношения правдоподобия.
- В случае, когда по отдельности какие-то фиктивные переменные не значимы, допустимо объединять уровни категориального предиктора, основываясь на интерпретации.
- Если какие-то уровни категориального предиктора лежат полностью в классе $y = 1$ или $y = 0$, их обязательно нужно объединить с другими уровнями, чтобы модель логистической регрессии могла быть построена.

Сравнение невложенных моделей

Невложенные модели можно сравнивать друг с другом по значению правдоподобия l, логарифма правдоподобия L или аномальности (deviance):

$$D = -2L.$$

Аномальность — аналог RSS в линейной регрессии; при добавлении признаков она не может убывать.

Для сравнения моделей с разным числом признаков можно использовать информационный критерий Акаике:

$$AIC = -2L + 2\left(k + 1\right) = D + 2\left(k + 1\right).$$

Мультиколлинеарность

Признаки мультиколлинеарности:

- правдоподобие модели высоко, но оценки многих коэффициентов близки к своим стандартным отклонениям;
- коэффициенты сильно меняются при включении и исключении других признаков.

Линейность логита

Проверка линейности логита по признакам — аналог визуального анализа остатков в обычной линейной регрессии.

Методы анализа линейности логита:

- сглаженные диаграммы рассеяния;
- фиктивные переменные по квартилям;
- дробные полиномы.

Сглаженные диаграммы рассеяния (smoothed scatterplots)

Рассмотрим оценку логита, полученную ядерным сглаживанием по x_j:

$$\bar{y}_{sm}(x_{ji}) = \frac{\sum\limits_{l=1}^{n} y_i K\left(\frac{x_{ji}-x_{li}}{h}\right)}{\sum\limits_{l=1}^{n} K\left(\frac{x_{ji}-x_{li}}{h}\right)},$$

$$\bar{l}_{sm}(x_{ji}) = \ln \frac{\bar{y}_{sm}(x_{ji})}{1-\bar{y}_{sm}(x_{ji})}.$$

График функции $\bar{l}_{sm}(x_j)$ должна быть похож на прямую.

Фиктивные переменные по квартилям (design variables)

1. Вычисляются выборочные квартили признака x_j: $[x_{j,0.25}, x_{j,0.5}, x_{j,0.75}]$.
2. Создаются три фиктивные переменные:

$$D_{1i} = [x_{j,0.25} \leq x_{ji} < x_{j,0.5}],$$
$$D_{2i} = [x_{j,0.5} \leq x_{ji} < x_{j,0.75}],$$
$$D_{3i} = [x_{j,0.75} \leq x_{ji}].$$

3. Настраивается логистическая регрессия, содержащая D_1, D_2 и D_3 вместо x_j. Пусть $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ — оценки коэффициентов при них.
4. Строится график, на котором середины интервалов, ограниченных квартилями, отложены против $\left[0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\right]$.

Дробные полиномы (fractional polynomials)

Если логит нелинеен по признаку, можно попробовать добавлять в модель его осмысленные степени и проверять их значимость.
В автоматическом режиме это можно делать с помощью дробных полиномов.

1. Настраиваются модели с заменой x_j на допустимые степени признака x_j, например, из множества $S = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$. Выбирается степень, максимизирующая правдоподобие.
2. Настраиваются модели с заменой x_j на двухкомпонентный полином x_j вида $\beta_{j_1} x_j^{p_1} + \beta_{j_2} x_j^{p_2}, \quad p_1, p_2 \in S$ (если $p_1 = p_2$, то берётся $\beta_{j_1} x_j^{p_1} + \beta_{j_2} x_j^{p_1} \ln x_j$). Выбираются степени, максимизирующая правдоподобие.
3. Если модель с полиномом второй степени значимо не лучше, чем линейная, используется линейная модель.
4. Если модель с полиномом второй степени значимо не лучше, чем с полиномом первой степени, используется модель с полиномом первой степени, иначе — с полиномом второй.

Содержательный отбор признаков

1. Если признаков достаточно много (например, больше 10), желательно сделать их предварительный отбор, основанный на значимости в однофакторной логистической регрессии. Для дальнейшего рассмотрения остаются признаки, достигаемый уровень значимости которых не превышает 0.25.
2. Строится многомерная модель, включающая все отобранные на шаге 1 признаки. Проверяется значимость каждого признака, удаляется небольшая группа незначимых признаков. Новая модель сравнивается со старой с помощью критерия отношения правдоподобия.
3. Чтобы убедиться, что удаление признаков не повлияло на оставшиеся, для каждого коэффициента $\hat{\beta}_j$ при оставшихся значимых признаках рассчитывается величина **delta-beta-hat-percent**:

$$\Delta\hat{\beta}\% = 100\frac{\hat{\beta}_j^{old} - \hat{\beta}_j^{new}}{\hat{\beta}_j^{new}}.$$

Если $\Delta\hat{\beta}\% > 20$, то какие-то из удалённых незначимых признаков были нужны, чтобы лучше определять коэффициенты значимых признаков; их нужно вернуть.

Содержательный отбор признаков

4. К признакам модели, полученной в результате циклического применения шагов 2 и 3, по одному добавляются удалённые признаки. Если какой-то из них становится значимым, он вносится обратно в модель.
5. Для непрерывных признаков полученной модели проверяется линейность логита. В случае обнаружения нелинейности признаки заменяются на соответствующие полиномы.
6. Исследуется возможность добавления в полученную модель взаимодействий факторов. Добавляются значимые интерпретируемые взаимодействия.
7. Проверяется адекватность финальной модели: близость y и $\hat{y}$; малость вклада наблюдений (x_i, y_i) на каждом объекте i в $\hat{y}$.

Порог классификации

Как по $\pi(x)$ оценить y?

$$y = [\pi(x) \geq p_0].$$

Чаще всего берут $p_0 = 0.5$, но можно выбирать по другим критериям, например, для достижения заданных показателей чувствительности или специфичности.

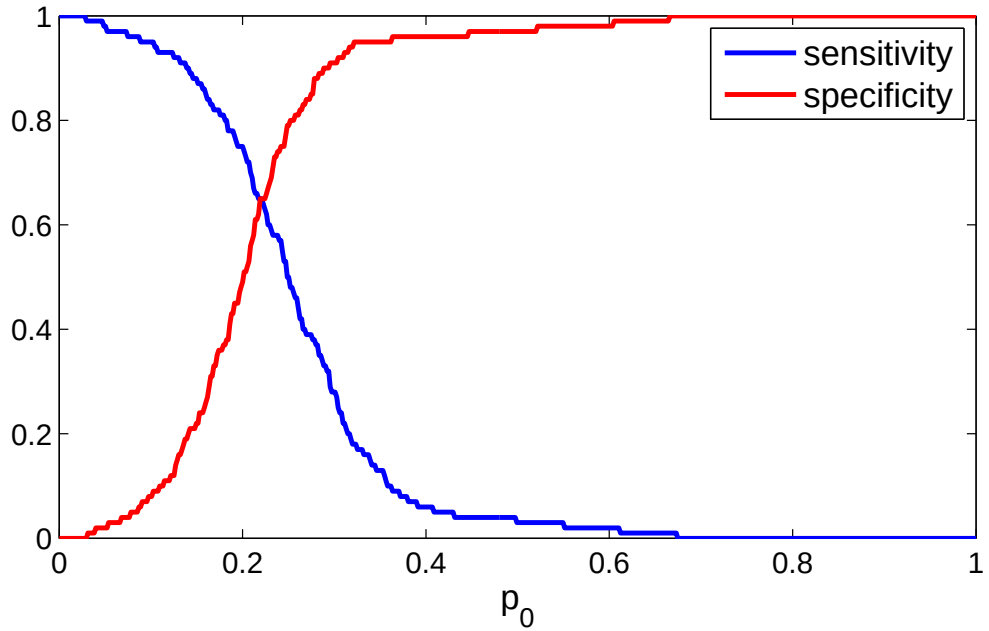
Порог классификации

Пример: эффективность терапии для наркозависимых, $p_0 = 0.5$:

$\hat{y}$ \ y	1	0
1	16	11
0	131	417

Чувствительность: $\frac{16}{16+131} \approx 10.9\%$.

Специфичность: $\frac{417}{11+417} \approx 97.4\%$.



Выбросы

Матрица проекции на пространство регрессоров (hat matrix):

$$H = V^{1/2} X \left(X^T V X \right)^{-1} X^T V^{1/2}.$$

Остатки Пирсона:

$$r_i = \frac{y_i - \hat{\pi}(x_i)}{\sqrt{\hat{\pi}(x_i)(1 - \hat{\pi}(x_i))}}.$$

Аналог расстояния Кука:

$$\Delta\hat{\beta}_i = \frac{r_i^2 h_i}{(1 - h_i)^2}.$$

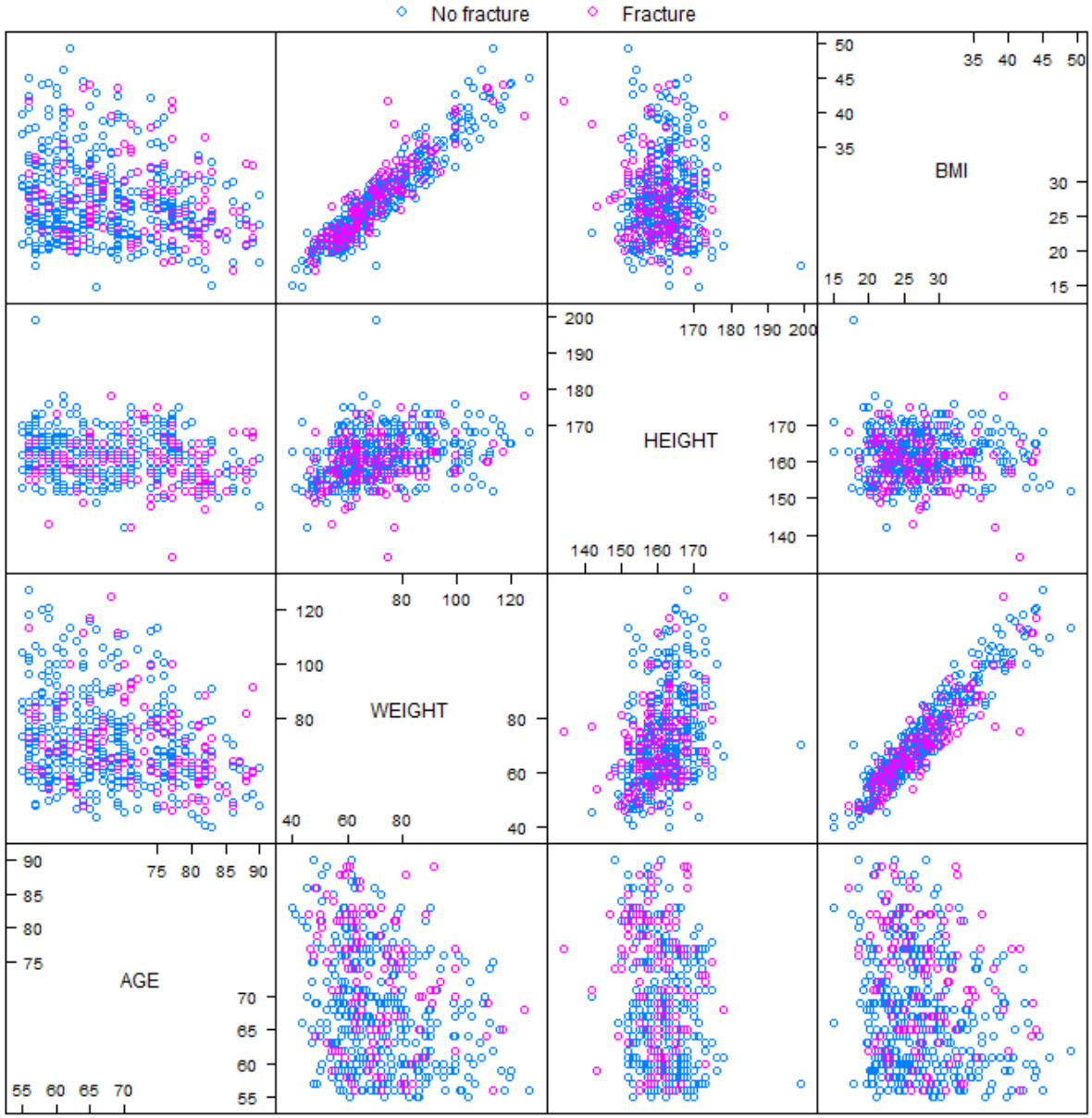
Риск остеопороза у женщин

Hosmer, Lemeshow, Sturdivant. Applied logistic regression: отобраны 500 участниц исследования Global Longitudinal Study of Osteoporosis in Women (Center for Outcomes Research, the University of Massachusetts/ Worcester). Измерены следующие показатели:

- возраст, лет (не меньше 55);
- вес, кг;
- рост, см;
- ИМТ, кг/см^2;
- бинарные признаки: курение, индикатор наступления менопаузы до 45 лет, индикатор необходимости помощи при подъёме из сидячего положения, перелом шейки бедра в прошлом (был/не было), перелом шейки бедра у матери (был/не было);
- самостоятельная субъективная оценка вероятности перелома (меньше/такая же/больше, чем у сверстниц).

Известно, у кого из участниц в первый год исследования произошёл перелом шейки бедра. Построить модель вероятности перелома с учётом имеющихся признаков.

Данные



Данные

$SMOKE$ \ y	0	1
no	347	118
yes	28	7

$PREMENO$ \ y	0	1
no	303	100
yes	75	25

$ARMASSIST$ \ y	0	1
no	250	62
yes	125	63

$MOMFRAC$ \ y	0	1
no	334	101
yes	41	24

$PRIORFRAC$ \ y	0	1
no	301	73
yes	74	52

$RATERISK$ \ y	0	1
1	139	28
2	138	48
3	98	49

Шаг 1

Результаты настройки одномерных моделей:

	coef	std	G	p
AGE	0.053	0.012	21.274	4.0×10^{-6}
WEIGHT	−0.005	0.006	0.665	0.4146
HEIGHT	−0.052	0.017	9.527	0.0020
BMI	0.006	0.017	0.112	0.7382
PREMENO	0.051	0.259	0.038	0.8451
PRIORFRAC	1.064	0.223	22.267	2.4×10^{-6}
MOMFRAC	0.661	0.281	5.270	0.0217
ARMASSIST	0.709	0.210	11.408	7.3×10^{-4}
SMOKE	−0.308	0.436	0.525	0.4687
RATERISK2	0.068	0.213	0.102	0.7489
RATERISK3	0.600	0.218	7.451	0.0006

Шаг 2

Результат настройки многомерной модели со всеми предикторами, значимыми на уровне не менее 0.25:

	coef	std	W	p
Intercept	2.709	3.230	0.839	0.4016
AGE	0.034	0.013	2.632	0.0085
HEIGHT	-0.044	0.018	-2.400	0.0164
PRIORFRAC	0.645	0.246	2.622	0.0087
MOMFRAC	0.621	0.307	2.024	0.0430
ARMASSIST	0.446	0.233	1.915	0.0555
RATERISK2	0.422	0.279	1.511	0.1307
RATERISK3	0.707	0.293	2.409	0.0160

Критерий отношения правдоподобия даёт:

- $p = 0.4535$ при сравнении модели шага 1 с текущей моделью;
- $p = 0.0508$ при оценке значимости признака RATERISK.

Шаг 3

$\Delta\hat{\beta}\%$:

kept \ deleted	WEIGHT	BMI	PREMENO	SMOKE
AGE	12.7	15.1	4.2	−4.2
HEIGHT	10.6	−3.6	−1.0	1.3
PRIORFRAC	−1.8	−1.9	−0.7	2.5
MOMFRAC	3.3	3.7	0.4	−0.1
ARMASSIST	−18.4	−19.7	−3.0	3.0
RATERISK2	0.5	0.7	−2.1	−0.9
RATERISK3	6.1	7.3	−0.7	−0.3

Порог в 20% не превышен $\Rightarrow$ ни один из удалённых признаков не оказывает существенного влияния на оценки оставшихся.

Шаг 4

Попробуем добавлять в полученную модель признаки WEIGHT, BMI, PREMENO и SMOKE; с помощью критерия отношения правдоподобия определим их значимость в получаемых моделях.

$$p_{WEIGHT} = 0.4131;$$
$$p_{BMI} = 0.3351;$$
$$p_{PREMENO} = 0.6703;$$
$$p_{SMOKE} = 0.4435.$$

Следовательно, признаки WEIGHT, BMI, PREMENO и SMOKE можно окончательно исключить.

Шаг 4

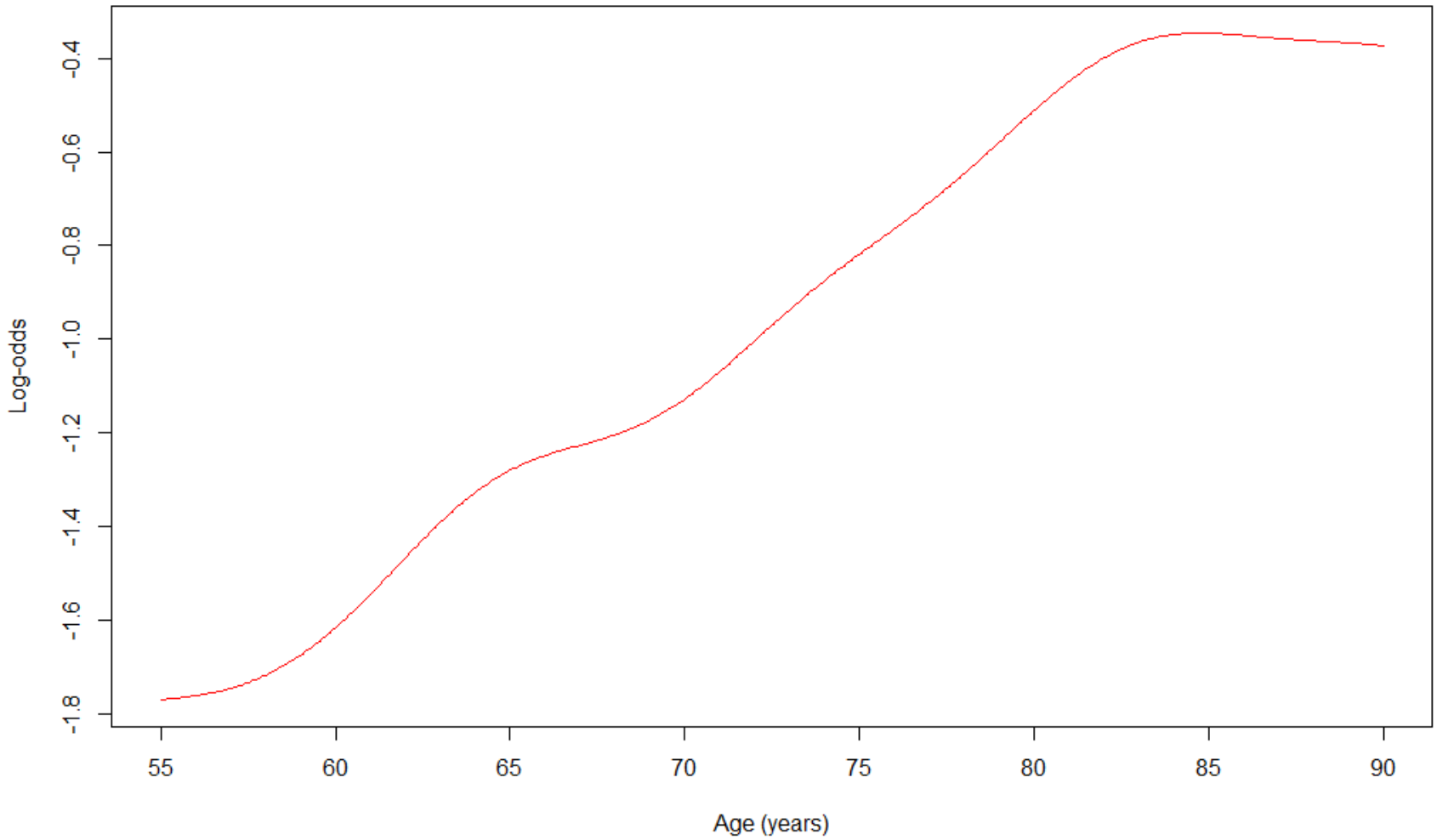
Переменная, кодирующая значение RATERISK=2, не значима (доверительный интервал для отношения шансов перелома при RATERISK=2 и в референсной категории RATERISK=1, содержит единицу). Возможно, эти два уровня стоит объединить (то есть, фактически, удалить фиктивную переменную RATERISK2). Специалисты, поставившие задачу, сочли такое преобразование осмысленным.

Новая модель:

	coef	std	W	p
Intercept	3.407	3.177	1.072	0.2836
AGE	0.033	0.013	2.563	0.0104
HEIGHT	−0.046	0.018	−2.555	0.0106
PRIORFRAC	0.664	0.245	2.709	0.0068
MOMFRAC	0.664	0.306	2.173	0.0298
ARMASSIST	0.473	0.231	2.044	0.0410
RATERISK3	0.458	0.238	1.924	0.0544

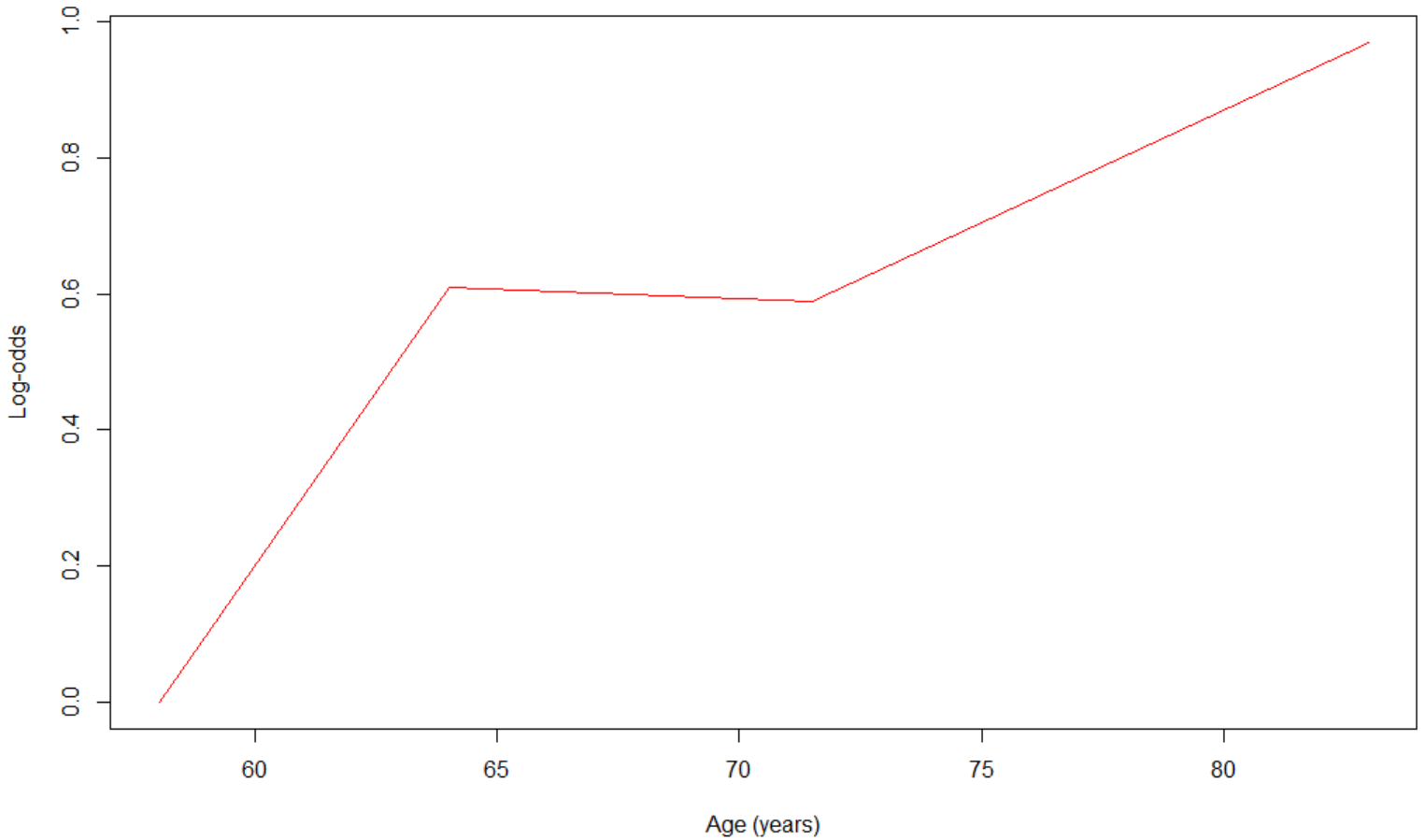
Шаг 5

Сглаженная диаграмма рассеяния:



Шаг 5

Фиктивные переменные по квартилям:



Квартиль	1	2	3	4
Диапазон	$x \leq 61$	$61 < x \leq 67$	$67 < x \leq 76$	$x > 76$
Середина	58	64	71.5	83
Коэффициент	0	0.610	0.590	0.970
Дов. инт.		$[-0.059, 1.278]$	$[-0.050, 1.229]$	$[0.311, 1.629]$

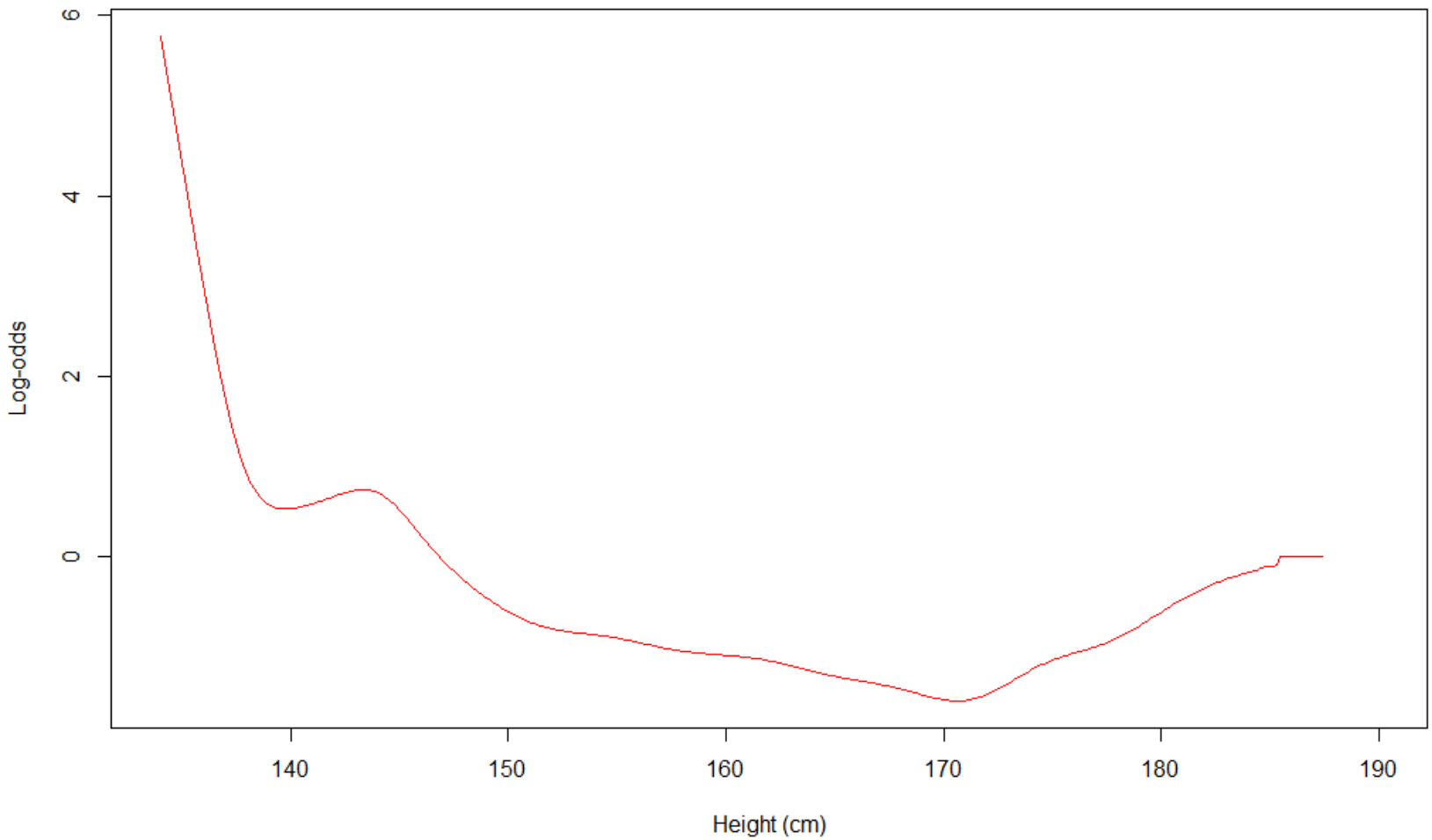
Шаг 5

Для модели, использующей вместо признака AGE признак $[AGE > 80]$, аномальность равна 512.8.
Аномальность модели, использующей признак AGE целиком, равна 509.8.

При переходе к бинаризованному признаку мы (всегда) теряем информацию и (в данном случае) не улучшаем модель $\Rightarrow$ не будем переходить к бинаризованному признаку.

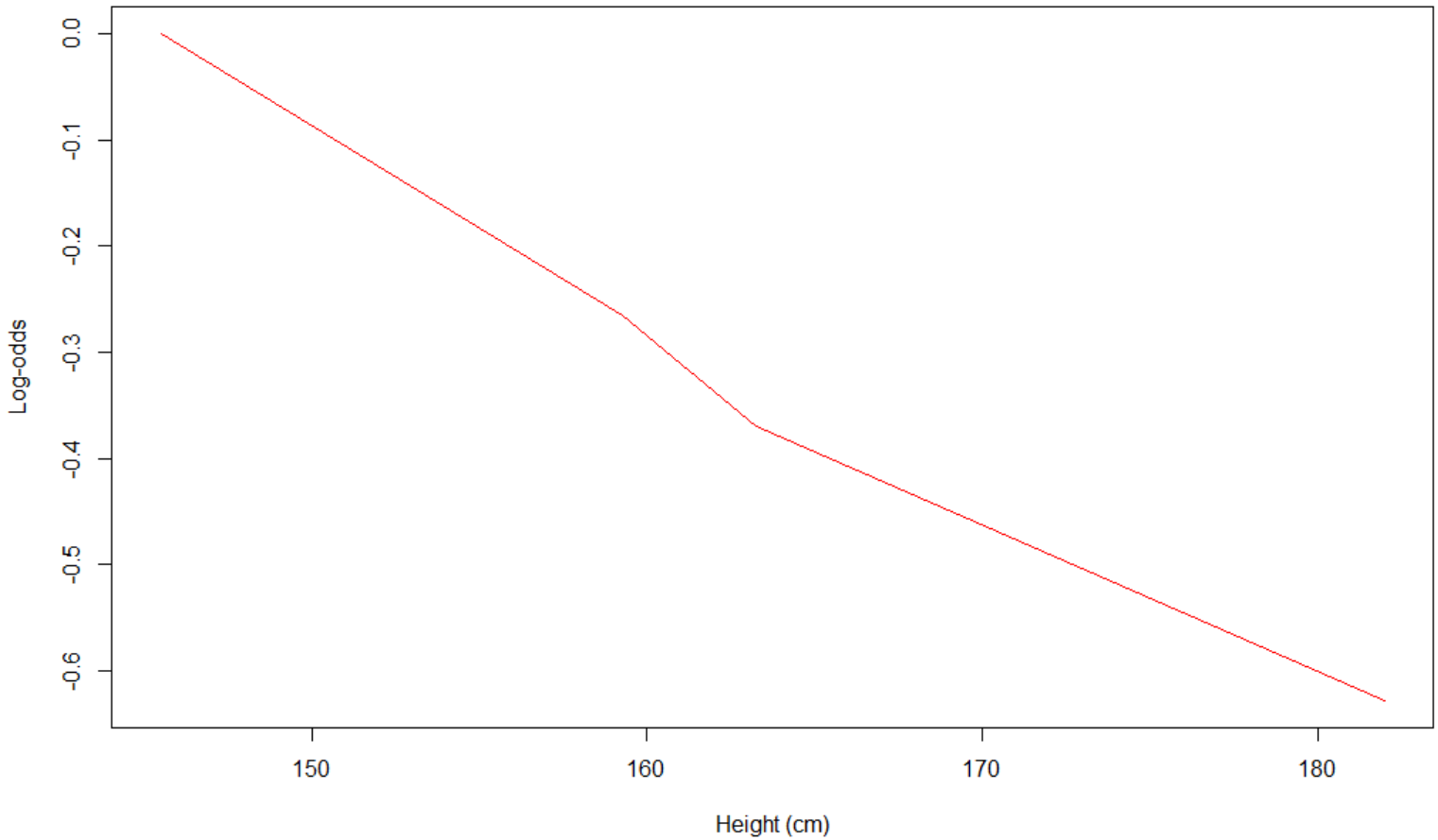
Шаг 5

HEIGHT, сглаженная диаграмма рассеяния:



Шаг 5

Фиктивные переменные по квартилям:



Квартиль	1	2	3	4
Диапазон	$x \leq 157$	$157 < x \leq 161.5$	$161.5 < x \leq 165$	$x > 165$
Середина	145.5	159.25	163.25	182
Коэффициент	0	-0.266	-0.369	-0.628
Дов. инт.		$[-0.861, 0.329]$	$[-0.964, 0.226]$	$[-1.255, -0.001]$

Шаг 5

Автоматический поиск не находит полиномов, позволяющих построить модель значимо лучше линейной:

```
> library(mfp)
> mfp(y ~ fp(AGE) + fp(HEIGHT) + PRIORFRAC + MOMFRAC + ARMASSIST +
   RATERISK3, family = binomial)

Fractional polynomials:
           df.initial select alpha df.final power1 power2
PRIORFRAC           1      1  0.05        1      1      .
HEIGHT              4      1  0.05        1      1      .
AGE                 4      1  0.05        1      1      .
MOMFRAC             1      1  0.05        1      1      .
ARMASSIST           1      1  0.05        1      1      .
RATERISK3           1      1  0.05        1      1      .

Transformations of covariates:
           formula
AGE          <NA>
HEIGHT       <NA>
PRIORFRAC    <NA>
MOMFRAC      <NA>
ARMASSIST    <NA>
RATERISK3    <NA>
```

Шаг 6

Рассмотрим все модели, в которые добавлено одно взаимодействие между факторами:

	LL	G	p
AGE*HEIGHT	-254.842	0.13	0.715
AGE*PRIORFRAC	-252.392	5.03	0.025
AGE*MOMFRAC	-254.840	0.14	0.710
AGE*ARMASSIST	-254.836	0.15	0.702
AGE*RATERISK3	-254.386	1.05	0.306
HEIGHT*PRIORFRAC	-254.802	0.21	0.645
HEIGHT*MOMFRAC	-253.704	2.41	0.121
HEIGHT*ARMASSIST	-254.111	1.60	0.207
HEIGHT*RATERISK3	-254.422	0.97	0.324
PRIORFRAC*MOMFRAC	-253.509	2.80	0.094
PRIORFRAC*ARMASSIST	-254.796	0.23	0.635
PRIORFRAC*RATERISK3	-254.848	0.12	0.726
MOMFRAC*ARMASSIST	-252.518	4.78	0.029
MOMFRAC*RATERISK3	-254.642	0.53	0.465
ARMASSIST*RATERISK3	-253.792	2.23	0.135

Шаг 6

Добавим все три взаимодействия, значимые на уровне 0.1:

	coef	std	W	p
Intercept	1.959	3.3272	0.59	0.556
AGE	0.058	0.0166	3.49	0.000
HEIGHT	-0.049	0.0184	-2.65	0.008
PRIORFRAC	4.598	1.8780	2.45	0.014
MOMFRAC	1.472	0.4229	3.48	0.000
ARMASSIST	0.626	0.2538	2.46	0.014
RATERISK3	0.474	0.2410	1.97	0.049
AGE*PRIORFRAC	-0.053	0.0259	-2.05	0.040
PRIORFRAC*MOMFRAC	-0.847	0.6475	-1.31	0.191
MOMFRAC*ARMASSIST	-1.167	0.6168	-1.89	0.058

Шаг 6

Уберём PRIORFRAC*MOMFRAC:

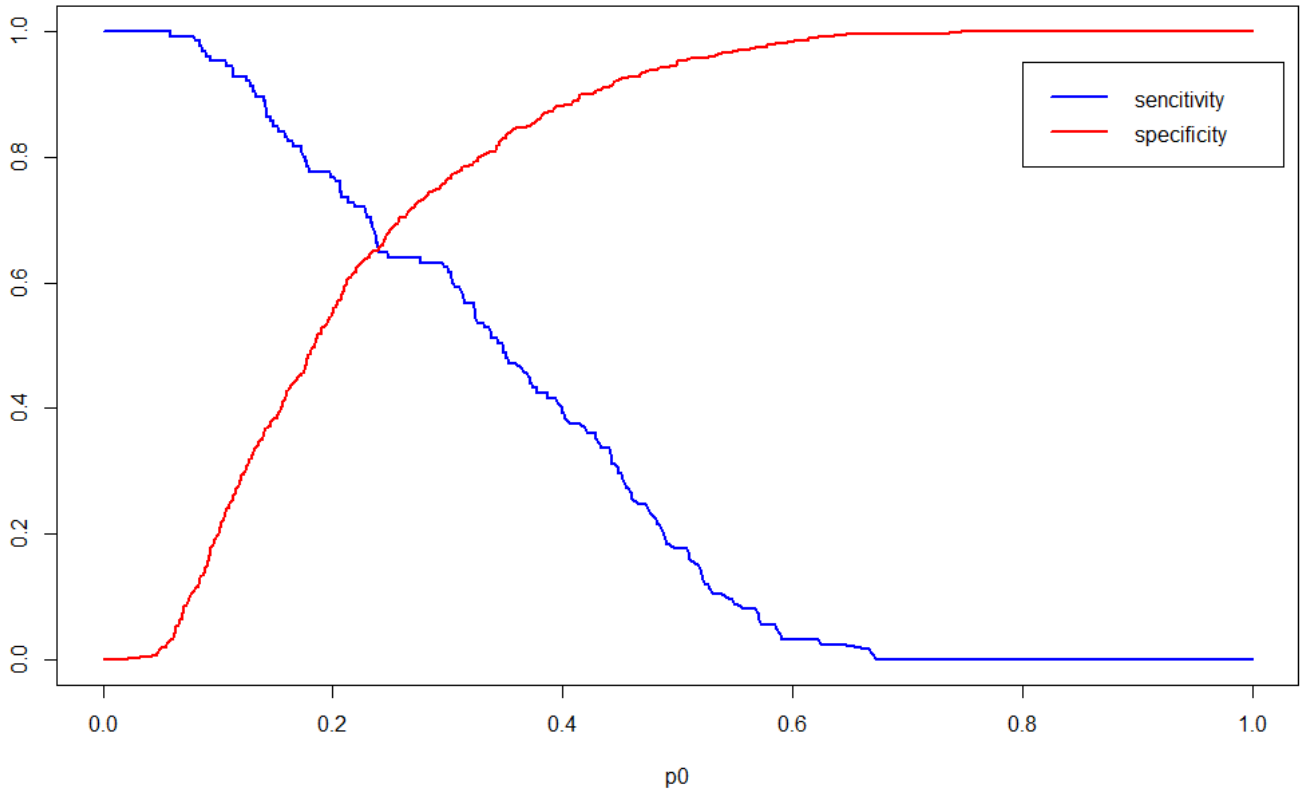
	coef	std	W	p
Intercept	1.717	3.3218	0.52	0.605
AGE	0.057	0.0165	3.47	0.001
HEIGHT	−0.047	0.0183	−2.55	0.011
PRIORFRAC	4.612	1.8802	2.45	0.014
MOMFRAC	1.247	0.3930	3.17	0.002
ARMASSIST	0.644	0.2519	2.56	0.011
RATERISK3	0.469	0.2408	1.95	0.051
AGE*PRIORFRAC	−0.055	0.0259	−2.13	0.033
MOMFRAC*ARMASSIST	−1.281	0.6230	−2.06	0.040

Порог классификации

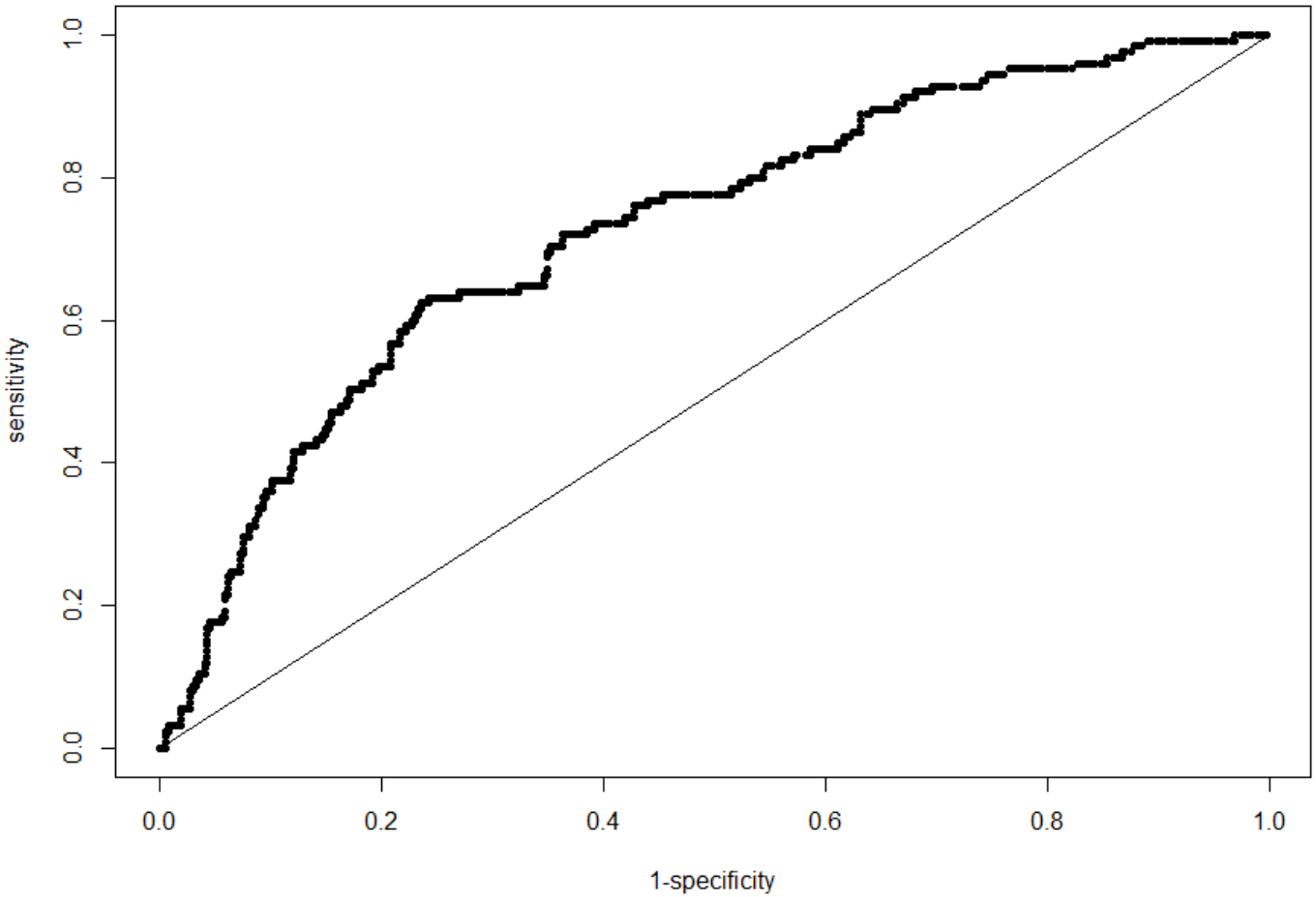
Результаты классификации с порогом 0.5:

$\hat{y}$ \ y	1	0
1	22	20
0	103	355

Чувствительность — $22/125 = 17.6\%$; специфичность — $355/375 = 94.7\%$.



Качество модели

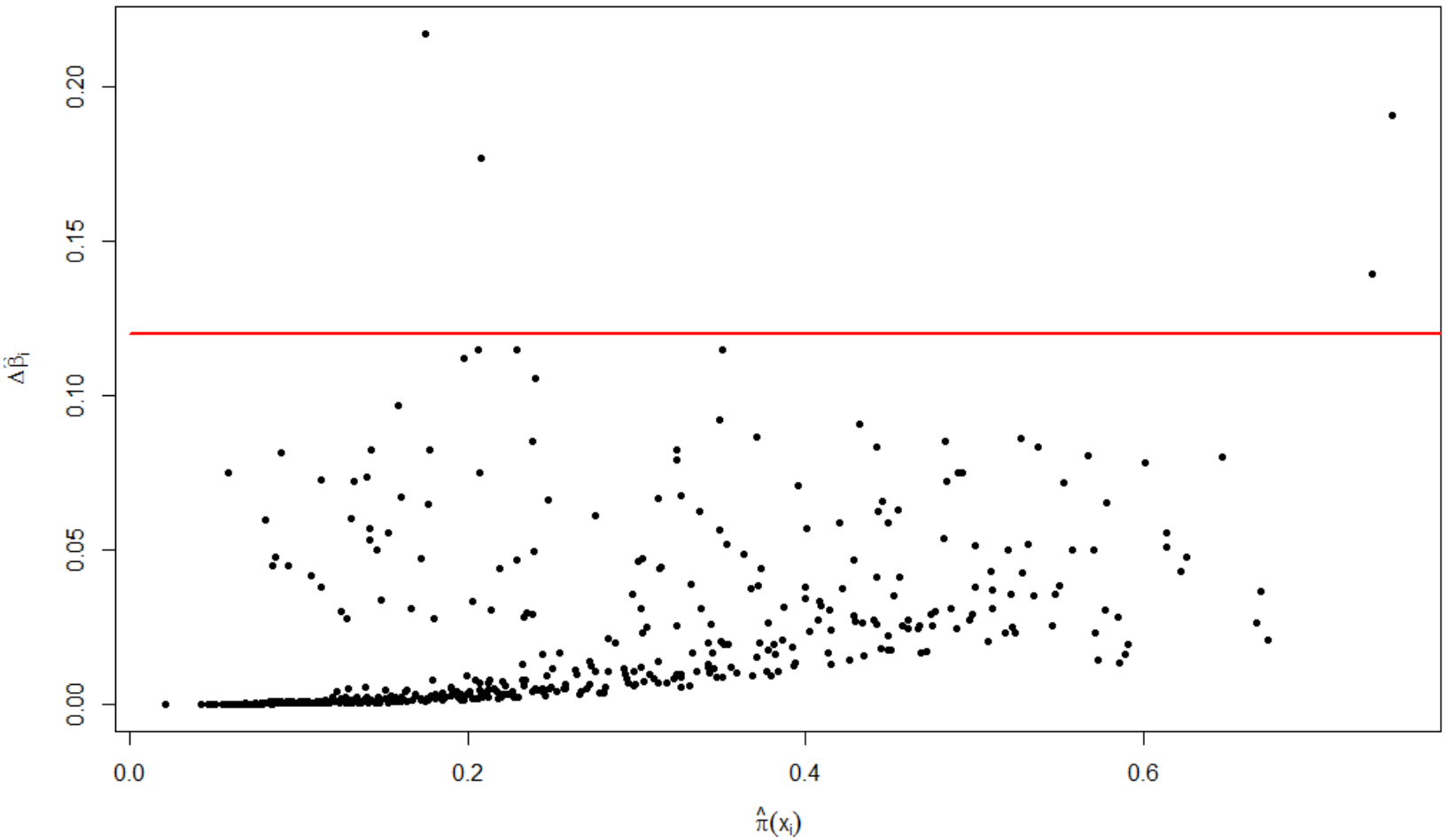


$AUC = 0.7286.$

Значимость модели по критерию отношения правдоподобия:
$p = 2 \times 10^{-10}$.

Качество модели

$\Delta\hat{\beta}_i$:



Качество модели

После удаления 4 наблюдений:

	coef	std	W	p
Intercept	3.570	3.436	1.04	0.2989
AGE	0.059	0.017	3.51	0.0005
HEIGHT	−0.059	0.019	−3.10	0.0019
PRIORFRAC	4.988	1.904	2.62	0.0088
MOMFRAC	1.505	0.404	3.73	0.0002
ARMASSIST	0.654	0.255	2.57	0.0103
RATERISK3	0.471	0.245	1.93	0.0541
AGE*PRIORFRAC	−0.059	0.026	−2.25	0.0243
MOMFRAC*ARMASSIST	−1.905	0.623	−2.87	0.0041

$\Delta\hat{\beta}_{MOMFRAC*ARMASSIST}\% = -32.8, \quad \Delta\hat{\beta}_{HEIGHT}\% = -21.3.$

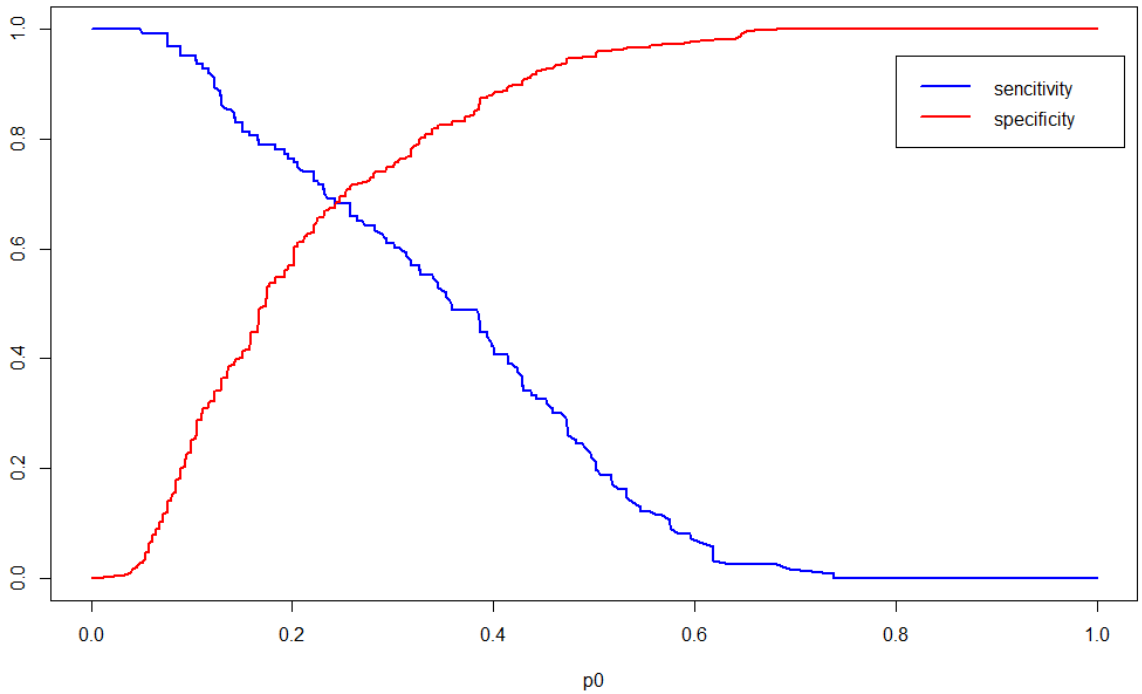
Значимость модели по критерию отношения правдоподобия:
$p = 8.9 \times 10^{-12}$.

Порог классификации

Результаты классификации с порогом 0.5:

$\hat{y}$ \ y	1	0
1	27	19
0	96	354

Чувствительность — $27/123 = 22.0\%$; специфичность — $354/373 = 94.9\%$.



$AUC = 0.7391$.

Результат

Итоговая модель вероятности перелома шейки бедра построена по 496 из 500 исходных наблюдений и определяет классификацию y с качеством $AUC = 0.7391$. При выборе порога по вероятности равным 0.5 модель обеспечивает классификацию с чувствительностью 22% и специфичностью 95%.

Модель позволяет сделать следующие выводы:

- для женщин, с которыми уже случался перелом шейки бедра, риск нового перелома в течение года существенно выше (в 146 раз, 95% доверительный интервал $(3.5, 6240)$);
- каждые десять лет относительный риск перелома шейки бедра у женщин, с которыми он ещё не происходил, возрастает в 1.8 раз (95% доверительный интервал $(1.3, 2.5)$), при этом для женщин, у которых уже был перелом, увеличение возраста не приносит значимого увеличения риска перелома (мультипликативный прирост риска за 10 лет 1.0001, 95% доверительный интервал $(0.4, 2.3)$);
- при прочих равных для женщин маленького роста риск перелома выше — на каждые 10 сантиметров уменьшения роста приходится увеличение риска перелома в 1.8 раз (95% доверительный интервал $(1.3, 2.7)$);

Результат

- для женщин, у матерей которых не было перелома шейки бедра, неспособность самостоятельно вставать из сидячего положения связана с повышением риска перелома в 1.9 раз (95% доверительный интервал $(1.2, 3.2)$);
- для женщин, которые способны самостоятельно вставать из сидячего положения, наличие перелома у матери связано с повышением риска перелома в 4.5 раз (95% доверительный интервал $(2.0, 9.9)$);
- женщины, высоко оценивающие вероятность перелома, действительно при прочих равных имеют шанс перелома выше в 1.6 раз, однако этот эффект слабо значим (95% доверительный интервал $(0.99, 2.6)$).

Требования к решению задачи методом логистической регрессии

- визуализация данных, оценка наличия выбросов, анализ таблиц сопряжённости по категориальным признакам;
- содержательный отбор признаков: выбор наилучшей линейной модели, оценка линейности непрерывных признаков по логиту, анализ необходимости добавления взаимодействий, провека адекватности финальной модели (классификация, анализ наличия выбросов);
- выводы.

Литература

Hosmer D.W., Lemeshow S., Sturdivant R.X. *Applied Logistic Regression*. — Hoboken: John Wiley & Sons, 2013.